Sistema de Procesamiento Digital de la Voz en el Dominio del Tiempo y la Frecuencia

Mat. Pablo Manrique Ramirez
Profesor e Investigador del CIC-IPN.
e-mail: pmanriq@vmredipn.ipn.mx

El sistema que se presenta inicializa con la digitalización de la señal eléctrica reproducida por un micrófono. La señal eléctrica es previamente transformada de una entrada analógica a una señal digital, para que posteriormente se realice el procesamiento sintético con el apoyo de un procesador DSP. Las técnicas digitales con el DSP facilitan sumamente el procesamiento de señales en tiempo real, lo cual no se podría realizar por técnicas analógicas.

El procesamiento digital con un DSP es muy seguro y puede ser llevado a cabo por un circuito compacto. El sistema se desarrollo en dos bloques de análisis, por criterios que están relacionados con el tipo de transformaciones por las que pasa la señal.

El primer bloque se encarga de controlar la señal desde un punto de vista analógico, para que posteriormente sea procesada por el segundo bloque, que se encarga de analizar la señal en forma digital. Así mismo, tiene la tarea de amplificar la señal en forma analógica a niveles manejables por el convertidor A/D.

El segundo bloque procesa la señal en forma digital, en base a algoritmos de análisis espectral, tales como la autocorrelación, la transformada rápida de Fourier (FFT), la transformada inversa rápida de Fourier (FFT^{-1}), el análisis homomorfológico y los filtros digitales.

Planteamiento del Sistema

Para lograr el entendimiento del desarrollo del sistema es necesario estudiar las características y factores del lenguaje hablado y de la señal que lo representa. La señal de la voz como una función del tiempo es una señal continua en tiempo, pues se presenta como un flujo de sonidos concatenados con o sin pausas entre ellos, lo que dificulta su segmentación en unidades simples (palabras, fonemas o sílabas) para su procesamiento. Lo ideal sería imitar el proceso de percepción y comprensión del ser humano. Pero esto es complicado, ya que para determinar la identidad de las palabras a partir de sus sonidos solamente, se hace necesario introducir más niveles de conocimiento del lenguaje, como por ejemplo: la acústica, la fonética, la lexicografía, la sintáctica, la semántica y la prosodia. En el presente sistema, sólo se consideran las siguientes características del lenguaje hablado como parte de la estrategia: se limita el universo de funcionamiento del sistema, es decir, se define un vocabulario limitado y una sintaxis rígida. Además, se enfocará para nuestros fines a los siguientes niveles de conocimiento del habla: acústica, fonética y prosodia, ya que éste se diseñó como un sistema práctico de laboratorio de procesamiento digital de señales.

Concepción del Diseño

El sistema contiene una interface en tiempo real donde, conforme a los análisis de las mediciones de los parámetros capturados, se entrega una serie de gráficas y resultados numéricos que representan el comportamiento de la señal de voz, que a la vez se procesa.

La **Figura 1** muestra un diagrama global a bloques del diseño del sistema de procesamiento de voz.

Figura 1. Diagrama global del sistema desarrollado.

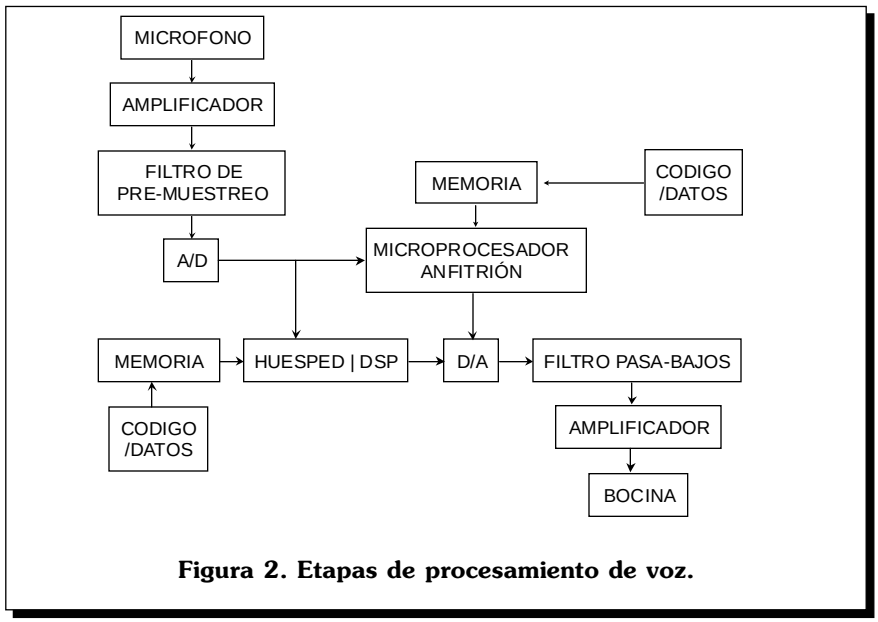

Figura 2. Etapas de procesamiento de voz.

Procesamiento Digital de la Señal de Voz

La **Figura 2** muestra un diagrama a bloques de las diferentes etapas por las que atraviesa la señal de voz, en el transcurso del sistema.

La parte inicial del sistema de preprocesamiento de la señal vocal está constituida por la señal analógica amplificada v(t) que se obtiene del micrófono. Antes de muestrear la señal de la voz, será necesario limitar la frecuencia máxima o ancho de banda frecuencial de ésta a la mitad de la frecuencia de muestreo, lo que se puede conseguir mediante un filtro analógico pasa-bajo previo al convertidor A/D, cuya frecuencia de corte sea la de Nyquist como máximo.

El ancho de banda frecuencial de la señal resultante deberá preservar la información necesaria para una adecuada descripción de los objetos acústicos. Así como es necesario el filtro pasa-bajas antes de la conversión A/D, es también necesario después de la conversión D/A para quitar las distorsiones presentes en las componentes armónicas altas.

El convertidor A/D realiza la tarea de muestrear y cuantificar la señal analógica v(t).

El proceso de muestreo consiste en convertir la señal analógica v(t), que se obtiene del bloque amplificador, en una secuencia de valores {v(kT)}, donde T es el período de muestreo.

Si representamos la señal de voz muestreada como $v_m(t)$, a intervalos de tiempo T=1/8KHz, pues el ancho de banda frecuencial de la señal de voz esta limitada a bajo de los 4 kHz, entonces :

$$v_m(t) = \sum_{k=-\infty}^{+\infty} v(kT)\ \delta(t-kT) \qquad (1)$$

Aplicando la Transformada Discreta de Fourier a la señal muestreada $v_m(t)$, obtenemos la expresión del espectro $V_m(j\omega)$, en función del espectro $V(j\omega)$ de la señal sin muetrear v(t). Suponiendo que la señal original v(t) en tiempo continua tiene un espectro limitado en banda, en el cual el radio de la componente en frecuencia más alta es ω_s, entonces, la expresión adopta la forma:

$$V_m(j\omega) = \frac{1}{T}\sum_{k=-\infty}^{+\infty} V[j(\omega+k\omega_m)];$$

$$\omega_m = 2\pi f_m \qquad (2)$$

Análisis Dependiente del Tiempo de la Señal de Voz

Para extraer los parámetros de la señal v(t), definimos la función o parámetro Γ de la siguiente forma:

$$\Gamma = \int_{-\infty}^{\infty} T[v(t)]dt \qquad (3)$$

en donde T es una transformación genérica invariente en el tiempo, y donde se debe observar que Γ no depende del tiempo, por lo tanto, representa una característica de la señal en la totalidad de su duración temporal. En el análisis de la señal v(t) nos interesa conocer la evolución temporal de las características de la señal, para esto se recurre a la aplicación en cada instante t de una «función ventana» n(t) que actúe como una función de peso y «realce» la señal transformada T[v(t)] en un determinado instante.

El parámetro Γ evaluado de esta manera vendrá determinado principalmente por las características instantáneas de la transformación de la señal v(t) alrededor de t. En este caso se obtiene la expresión:

$$\Gamma(t) = \int_{-\infty}^{\infty} T[v(t)]\ v(\tau-t)\ d\tau \qquad (4)$$

que define el valor del parámetro Γ en cada instante t. Debe observarse que esta expresión tiene exactamente la forma de una convolusión entre la señal transformada T[v(t)] y la ventana n(t), por lo que se puede interpretar Γ(t) como la salida de un filtro lineal pasa-bajo, cuya respuesta al impulso es n(t), y cuya señal de entrada es T[v(t)].

El ancho de banda del filtro representado por la ventana es inversamente proporcional a la longitud temporal efectiva de su respuesta al impulso n(t). La longitud de la ventana tiene una influencia fundamental sobre la función-parámetro Γ(t), cuya máxima rapidez de variación vendrá determinada por el ancho de banda del filtro equivalente a la ventana. Cuanto más estrecha sea dicha banda (más larga será la ventana temporal n(t)) y, por lo tanto, menos se reflejarán en Γ(t) las variaciones rápidas de v(t). La longitud de la ventana deberá elegirse, entonces, en función del suavizado o promediado que se desee para Γ(t); una ventana muy larga promediara Γ(t) en un intervalo demasiado grande, perdiendo las variaciones locales; una ventana muy corta reflejará excesivamente sobre Γ(t) las fluctuaciones instantáneas de v(t). Por lo tanto, para una determinada duración de la ventana, la forma de la función n(t) escogida determinará el perfil del filtro, lo cual afectará al parámetro resultante Γ(t) en mayor o menor grado según el tipo de transformación T utilizada en la definición de Γ(t).

Análisis en el Dominio del Tiempo

La expresión de la energía para señales muestreadas es:

$$E[n]=\sum_{k=-\infty}^{+\infty} v^2[m]\ v[m-n] \qquad (5)$$

Para el reconocimiento de una palabra, la energía proporciona una primera aproximación para distinguir segmentos vocálicos (alta energía) de segmentos consonánticos (baja energía) y, en el caso de señal de buena calidad (alta relación señal/ruido), se le puede utilizar para distinguir la voz del silencio (detección de bordes). Pero para nuestros fines el uso de la energía como parámetro puede representar el inconveniente de su gran sensibilidad a la amplitud de la señal (ya que ésta aparece elevada al cuadrado), que implica la necesidad de un gran margen dinámico o el uso de una transformación logarítmica. Otro inconveniente lo constituye la complejidad de cálculo que supone la evaluación de los cuadrados. Para evitar estas dificultades se suele utilizar alternativamente la amplitud media dependiente del tiempo, definida como:

$$A[n]=\sum_{k=-\infty}^{+\infty} |v[m]|\ v[m-n] \qquad (6)$$

La señal v[m] produce un cruce por cero cuando cambia de signo. Definimos la función sig(v[k]) como sigue:

$$sig(v[k])=\begin{cases} 1, & v[k]>0 \\ 0, & v[k]\le 0 \end{cases}$$

Entonces se calcula la densidad de cruces por cero dependiente del tiempo de la señal en su forma discreta como sigue:

$$Z[n]=\frac{1}{L}\sum_{k=-\infty}^{+\infty} |sig(v[m]) - sig(v[m-1])|\ v[m-n] \qquad (7)$$

donde L representa la duración efectiva de la función ventana n[m] utilizada. Este es un parámetro de muy baja complejidad de cálculo, y se le ha utilizado para detectar segmentos fricativos (señal de pequeña energía y elevada densidad de cruces por cero).

Análisis en el Dominio de la Frecuencia

Anteriormente consideramos algunos casos simples de análisis dependiente del tiempo, en los que la transformación T aplicada a la señal de la voz, conduce a parámetros Γ(t) de tipo escalar. Existen transformaciones de la señal en las que la función resultante G(t) es, a su vez, función de otra variable Z(Γ(z,t)). Entre las transformaciones de este tipo cabe señalar, tanto por su utilidad práctica como conceptual en análisis de la señal vocal, la transformada de Fourier (TF). En varias aplicaciones de procesamiento de señales, incluyendo la voz, es imposible satisfacer las condiciones de la TF y es imparcial especificar la señal para todo tiempo. Todos estos problemas pueden ser superados si la señal muestreada de la ecuación (1) se trunca y una porción estacionaria de esta se usa en la computación espectral. Supóngase que esta porción de la señal consiste de N muestras y sustituyase en la TF, es decir:

$$V(\omega)=\int_{-\infty}^{\infty}\left\{\sum_{n=0}^{N-1}\delta(t-nT)v(nt)\right\}e^{-i\omega t}dt \qquad (8)$$

luego entonces, el espectro en frecuencia es periódico y se repite en múltiplos enteros de la frecuencia de muestreo $w_m = 2p/T$. Para los propósitos de computación, se asume que el espectro es evaluado en N frecuencias, linealmente espaciadas entre el centro del espectro y la frecuencia de muestreo w_m. Si la frecuencia entre componentes de frecuencia sucesivas es dw entonces:

$N\delta\omega=\omega_m$ y $\omega=k\delta\omega$, k=0,1,2,...,N-1

$$V(k\delta\omega)=\int_{-\infty}^{\infty}\left\{\sum_{n=0}^{N-1}\delta(t-nT)v(nt)\right\}x\ e^{-ik\delta\omega t}dt \qquad (9)$$

Intercambiando el orden de la integración y la suma, y tomando en cuenta que:

$dw=w_m/N$ y $w_m=2p/T$,

por lo tanto dw=2p/TN;

la ecuación anterior puede escribirse:

$$V(k\delta\omega)=\sum_{n=0}^{N-1} v(nt)e^{-j2\pi kn/N}; \quad (10)$$

k=0,1,2,...,N-1

la expresión anterior se conoce como la transformada discreta de Fourier (DFT) de la señal de voz. Una expresión para la transformada discreta inversa de Fourier (DFT^{-1}) se deriva similarmente:

$$v(nT)=\frac{1}{N}\sum_{n=0}^{N-1} V(k\delta\omega)e^{j2\pi kn/N}; \quad (11)$$

k=0,1,2,...,N-1

El cómputo de la DFT para una sucesión de N-puntos produce una serie de términos complejos, que contienen la amplitud y la fase de información para cada componente espectral, y se lleva acabo a través del algoritmo de la FFT.

Análisis Ceptral de Voz

Para la extracción de la envolvente del tracto vocal, se requiere una técnica para quitar los rizos fundamentales. Esto se lleva a cabo por una técnica conocida como truncación ceptral, la cual ahora se describirá.

Sea V(w) el espectro del sonido de la señal de voz, P(w) el espectro de los impulsos fundamentales y H(w) el espectro del tracto vocal, el cual incluye los efectos de configuración de la forma de onda glotal y la radiación desde la boca. La relación entre la magnitud de estos tres espectros puede ser expresada simplemente como sigue:

$$|V(\omega)| = |P(\omega)| \times |H(\omega)| \quad (12)$$

Obteniendo el logaritmo de la ecuación precedente, se obtiene:

$$\log|V(\omega)| = \log|P(\omega)| + \log|H(\omega)| \quad (13)$$

Esto conduce a la anotación de que estas dos componentes deben ser separables por medio de una operación de filtro lineal. Este filtración es normalmente transportado por la transformada inversa de Fourier de log½V(w)½ para producir lo que se conoce como el cepstrum de la señal. El eje horizontal del cepstrum tiene dimensión en tiempo y está condicionado a la frecuencia de la señal de la voz. Es claro que la truncación cepstral es un método extremadamente útil para quitar «tono de murmullo» desde un espectro de alta resolución, para dejar sola la información de la función de transferencia del tracto vocal.

Conclusiones

Las técnicas utilizadas permiten analizar la señal de la voz, contribuyen al mejoramiento y manipulación de sus características en el dominio del tiempo y en el dominio de la frecuencia, donde se puede destacar que son más efectivas las técnicas espectrales.

Bibliografia

[1] F. J. Owens. *"Signal Processing of Speech"*. McGraw-Hill, Inc.

[2] Lawrence. R. Rabiner, Ronald. W. Schafer. *"Digital Processing of Speech Signals"*. Prentice-Hall.

[3] Sadaoki Furui.*"Digital Speech Processing, Synthesis, and Recognition"*. Marcel Dekker, Inc.

[4] Lawrence Rabiner, Biing-Hwang Juang. *"Fundamentals of Speech Recognition"*. Prentice-Hall.